Construction d'un score d'événement à court terme pour les insuffisants cardiaques

Kévin Duarte [1,2,3,*], Jean-Marie Monnez [1,2,3,4,†] & Eliane Albuisson [1,5,6,‡]

[1] *Université de Lorraine, CNRS, Institut Elie Cartan de Lorraine, F-54000 Nancy, France*

[2] *INRIA, Project-Team BIGS, F-54600 Villers-lès-Nancy, France*

[3] *INSERM U1116, Centre d'Investigations Cliniques-Plurithématique 1433, Université de Lorraine, Nancy, France*

[4] *Université de Lorraine, Institut Universitaire de Technologie Nancy-Charlemagne, F-54052 Nancy, France*

[5] *BIOBASE, Pôle S2R, CHRU de Nancy, Vandoeuvre-lès-Nancy, France*

[6] *Faculté de Médecine, InSciDenS, Vandoeuvre-lès-Nancy, France*

∗ k.duarte@chru-nancy.fr ; † jean-marie.monnez@univ-lorraine.fr ;

‡ e.albuisson@univ-lorraine.fr

Financement : Programme Investissements d'Avenir ANR-15-RHU-0004

Résumé. L'insuffisance cardiaque (IC) est un problème majeur de santé publique. Afin d'identifier les patients à risque de décéder ou d'être hospitalisé pour progression de l'IC à court terme, nous avons construit un score d'événement d'IC par l'intermédiaire d'une méthode d'ensemble, en utilisant deux règles de classification différentes, la régression logistique et l'analyse discriminante linéaire de données mixtes, des échantillons bootstrap, et en introduisant un aléa dans la construction des prédicteurs par une sélection aléatoire de variables. L'intervalle de variation du score a été ramené sur une échelle de 0 à 100. Enfin, nous définissons une mesure du risque d'événement associé au score par un odds-ratio et mesurons l'importance des variables et des groupes de variables en utilisant les coefficients standardisés.

Mots-clés. *Apprentissage et classification, biostatistique, méthode d'ensemble, score d'événement, insuffisance cardiaque.*

Abstract. Heart failure is a major public health problem. In order to identify patients at risk of death or hospitalizion for progression of HF in the short term, we built HF event score via an ensemble method using two classification rules, logistic regression and linear discriminant analysis, boostrap samples and introducing a randomness into the construction of models. The range of variation of score was rescaled from 0 to 100. Finally we defined a measure of event risk by an odds ratio and measured the importance of variables or groups of variables using standardized coefficients.

Keywords. *Learning and classification, biostatistics, ensemble method, event score, heart failure.*

1 Introduction

L'insuffisance cardiaque (IC) est un problème majeur de santé publique qui contribue de façon importante à la mortalité par maladies cardiovasculaires. Afin d'identifier les patients à risque de décéder ou d'être hospitalisé pour progression de l'IC à court terme, nous nous sommes intéressés au problème de la construction d'un score d'événement d'IC à court terme à partir de mesures cliniques, démographiques, biologiques et des antécédents médicaux d'un patient.

2 Données

Notre étude a été réalisée sur des données provenant de l'étude clinique internationale multicentrique EPHESUS qui incluait 6632 patients atteints d'insuffisance cardiaque (IC) aiguë après un infarctus du myocarde avec une dysfonction systolique du ventricule gauche, dont les résultats sont présentés dans Pitt, Remme, Zannad et al. (2003). Au cours de cette étude, des visites de suivi pour chaque patient ont été réalisées à l'inclusion du patient dans l'étude, 1 mois après l'inclusion, 3 mois après, puis tous les trois mois jusqu'à la fin du suivi. A chaque visite, de nombreux paramètres biologiques, cliniques ou d'antécédents médicaux ont été observés et tous les événements indésirables (décès, hospitalisations, maladies) survenus au cours du suivi ont été collectés.
A partir des mesures biologiques, cliniques ou d'antécédents médicaux observées sur un patient à un temps fixé, nous cherchons à évaluer le risque que ce patient ait un événement d'IC à court terme. Les individus que nous considérons sont des couples (patient-temps). Nous supposons donc que le devenir à court terme du patient ne dépend que de ses mesures actuelles.
La variable à expliquer est la survenue ou non d'un événement composite d'IC à court terme (décès ou hospitalisation pour progression de l'IC). Afin de disposer de suffisamment d'événements, nous avons défini le court terme à 30 jours. Ainsi, les individus considérés sont des patients-mois. Après examen complet de la base de données, nous avons pu définir un ensemble de 27 variables explicatives candidates, réparties en trois catégories : données cliniques, mesures biologiques et historique médical.
Nous disposons au final de 21382 observations provenant de 5937 patients différents, dont 317 observations avec événement d'IC à 30 jours.

3 Construction du score

3.1 Méthodologie générale

Pour construire le score, nous utilisons une méthode d'ensemble qui peut être présentée sous la forme d'un arbre (cf. figure 1). On peut trouver un exposé sur les méthodes

d'ensemble dans Genuer et Poggi (2017) et un exemple de méthode d'ensemble, RGLM (Random Generalized Linear Model), dans Song, Langfelder et Horvath (2013). Nous mesurons la qualité d'un prédicteur par l'AUC.

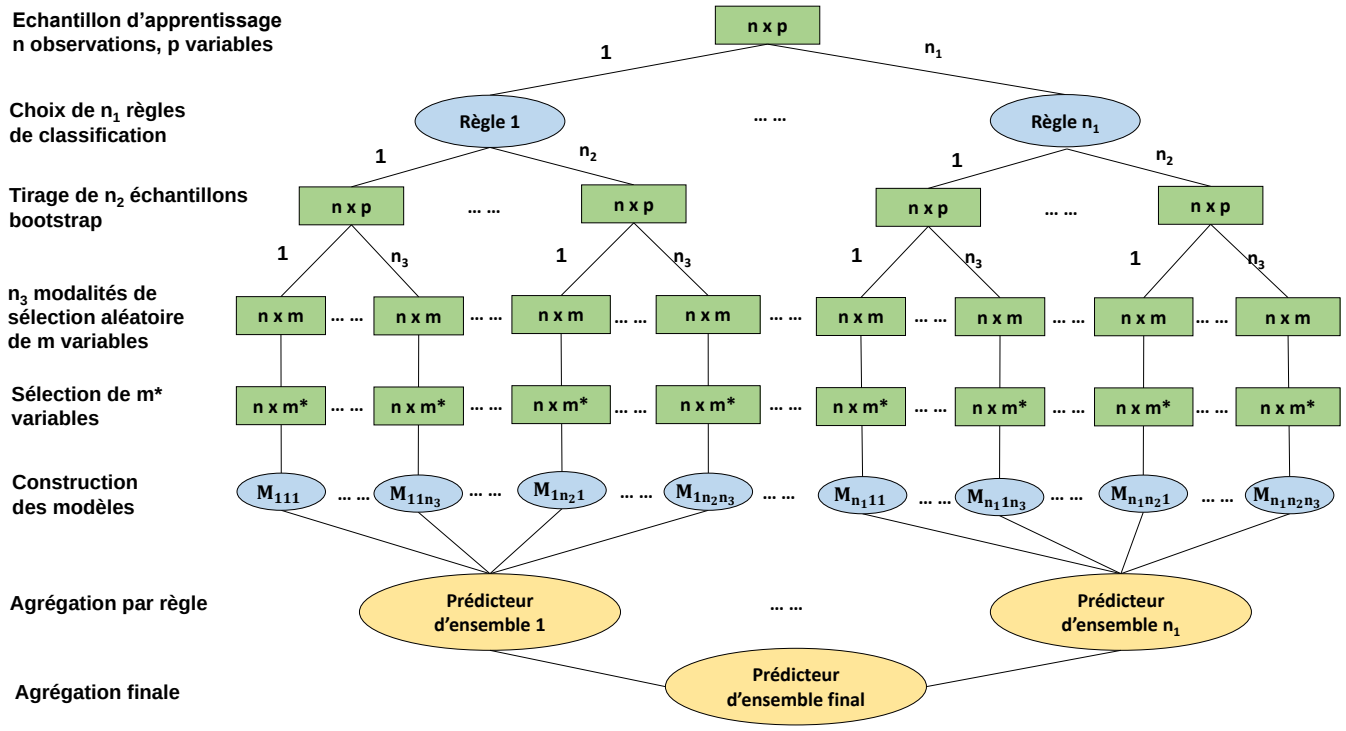

FIGURE 1 – Méthodologie de construction d'un score

Au 1er niveau de l'arbre, on retient n_1 règles de classification. Après essai de différentes méthodes, nous avons retenu la régression logistique et l'analyse discriminante linéaire (LDA).
Au 2e niveau de l'arbre, pour chaque règle de classification, on génère n_2 échantillons bootstrap ($n_2 = 1000$).
Au 3e niveau de l'arbre, on retient n_3 modalités de sélection aléatoire de variables, une modalité étant définie soit par un nombre de variables tirées au hasard, soit par un nombre de groupes prédéfinis de variables corrélées, tirés au hasard, à l'intérieur de chacun desquels on tire au hasard une variable. Nous avons retenu $n_3 = 3$ modalités. Le nombre de variables ou de groupes de variables tirés est celui qui maximise l'AUC.
Au 4e niveau de l'arbre, on retient une seule modalité de sélection de variables, par une méthode stepwise ou de pénalisation (LASSO, ridge ou elastic net). Pour notre problématique, il est apparu que ce niveau n'apportait pas d'amélioration de la qualité de prédiction, d'après les expériences effectuées. Nous ne l'avons donc pas retenu.

Au 5e niveau de l'arbre, on construit les modèles selon les modalités définies aux niveaux précédents. Pour chaque règle de classification, ceci donne un ensemble de $n_2 \times n_3$ prédicteurs.
Au 6e niveau de l'arbre, on réalise une première agrégation par règle de classification.
En LDA, on utilise comme fonction-score la fonction linéaire discriminante de Fisher

$$S_1(x) = \left(x - \frac{g_1 + g_0}{2}\right)' M(g_1 - g_0) = \alpha_1' x + \beta_1.$$

En régression logistique, on peut utiliser comme fonction-score

$$S_2(x) = ln\frac{P(\Omega_1|X = x)}{P(\Omega_0|X = x)} = \alpha_2' x + \beta_2.$$

Pour chaque règle de classification, nous avons fait la moyenne des fonctions-scores obtenues. Ainsi, on obtient deux fonctions-scores synthétiques : une par LDA notée $\bar{S}_1$ et une autre par régression logistique notée $\bar{S}_2$.
Au 7e niveau de l'arbre, on réalise une agrégation des deux scores synthétiques $\bar{S}_1$ et $\bar{S}_2$ construits à l'étape précédente. Pour cela, nous avons considéré une combinaison $(1 - \lambda)\bar{S}_1 + \lambda\bar{S}_2$ et cherché λ compris entre 0 et 1 qui maximise l'AUC. L'AUC calculée en resubstitution étant généralement trop optimiste, nous avons évalué la capacité de généralisation du score par l'AUC OOB.
Nous avons ramené l'intervalle de variation de la fonction-score synthétique de 0 à 100. Ce score donne un AUC en resubstitution de 0.8733 et un AUC OOB de 0.8667.

3.2 Mesure du risque par un odds-ratio

On pourra interpréter ce score en calculant une mesure du risque associé à un score s par un odds-ratio. Nous en donnons deux définitions :

$$\begin{aligned} OR_1(s) &= \frac{P(Y = 1|S > s)}{P(Y = 0|S > s)} \times \frac{P(Y = 0)}{P(Y = 1)} = \frac{P(S > s|Y = 1)}{P(S > s|Y = 0)} = \frac{Se(s)}{1 - Sp(s)} \\ OR_2(s) &= \max_{t \leq s : OR_1(t) < \infty} OR_1(t) \end{aligned}$$

OR_2 permet d'effectuer un lissage de OR_1. L'évolution des odds-ratio OR_1 et OR_2 est représentée sur la figure 2.

3.3 Importance des variables

Les coefficients bruts des variables dans le score ne sont pas de bons indicateurs de leur importance, car ces variables n'ont pas la même unité. Ainsi, nous avons calculé les coefficients "standardisés", en multipliant chaque coefficient obtenu par l'écart-type de la

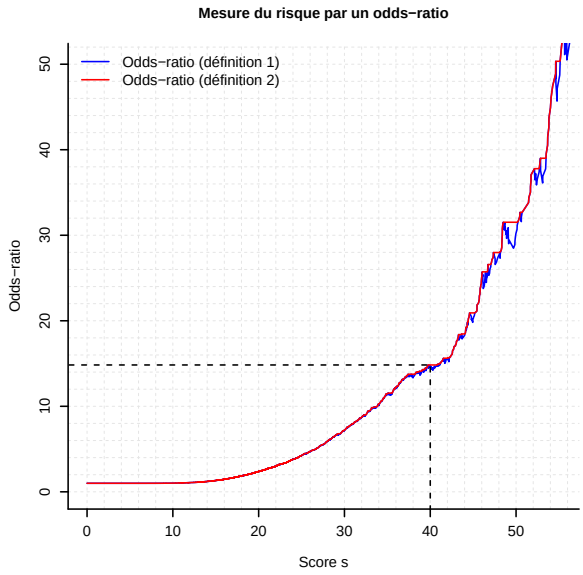

FIGURE 2 – Mesure du risque par un odds-ratio

variable correspondante. Le fait de standardiser les coefficients permet de ne pas tenir compte des unités et les coefficients deviennent directement comparables. Pour avoir une vision globale de l'importance des variables dans le score, nous avons représenté sur un graphique pour chaque variable la valeur absolue de son coefficient standardisé, de la plus grande valeur à la plus petite (cf. figure 3). Le même type de graphique a été réalisé pour représenter l'importance des groupes de variables corrélées que nous avons constitués. Cette fois-ci, on représente pour chaque groupe de variables la somme des valeurs absolues des coefficients standardisés associés aux variable du groupe, de la plus grande somme à la plus petite (cf. figure 3).

4 Apprentissage en ligne

Dans cette étude, nous avons présenté une méthodologie de construction d'un score d'IC à court terme, basée sur la construction d'un prédicteur d'ensemble. Cette méthode peut être mise en œuvre dans le cadre de l'apprentissage en ligne, en utilisant des algorithmes de gradient stochastique pour mettre à jour en ligne les prédicteurs. Dans le cas de la régression linéaire, et en particulier de l'analyse discriminante linéaire binaire, plusieurs processus ont été étudiés et comparés entre eux et à plusieurs processus classiques dans Duarte, Monnez et Albuisson (2018). Pour éviter les problèmes d'explosion numérique, nous avons utilisé des données standardisées en ligne, en estimant récursivement les moyennes et les écarts-types des variables parallèlement aux processus de gradient stochastique. Dans le

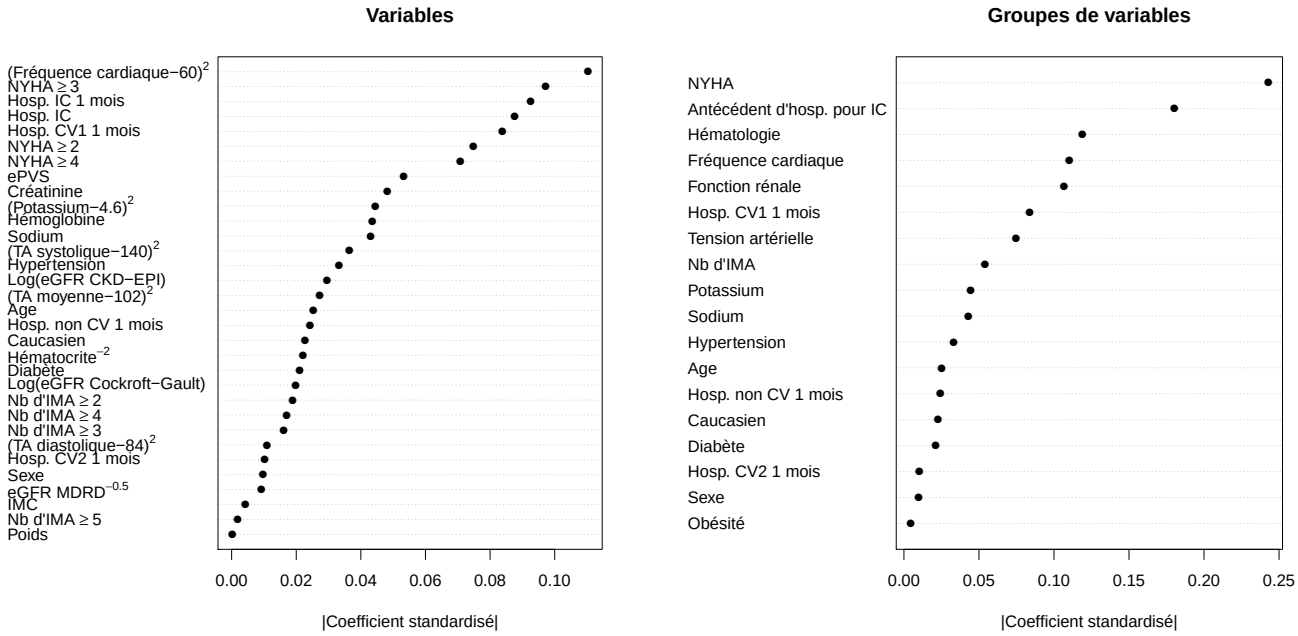

FIGURE 3 – Importance des variables et des groupes de variables

cas de la régression logistique, la convergence d'un processus avec données standardisées en ligne a été établie par Monnez (2017).

Bibliographie

[1] Pitt, B., Remme, W., Zannad, F. et al. (2003), Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction, N Engl J Med, 348 :1309-1321.
[2] Genuer, R. and Poggi, J.M. (2017). Arbres CART et Forêts aléatoires, Importance et sélection de variables, Working paper or preprint.
[3] L. Song, P. Langfelder, and S. Horvath (2013), Random generalized linear model : a highly accurate and interpretable ensemble predictor, BMC bioinformatics, 14(1) :5.
[4] Duarte, K., Monnez, J.M. and Albuisson, E (2018), Sequential linear regression with online standardized data, PloS One, 13 (1) e0191186.
[5] Monnez, J.M. (2017), Sequential logistic regression with online standardized data, Working paper.